COP328

MSc Data Science Project Research Proposal

Predicting Customer Life-time Value for eCommerce Platforms

ID Number: B614677

May 5, 2021

Supervisor: Eran Edirisinghe

Examiners: Dr Georgina Cosma
Professore Sara Lombardo

Loughborough University
Computer Science Department
LE11 2TU, Epinal Way
Loughborough, Leicestershire

1 Introduction

1.1 Project Focus & Background

Ecommerce businesses invest in their customers by deciding on acquisition costs, advertisements, promotions, product discounts, etc. in order to generate revenue and be profitable. There are always some customers who lower this profitability however, while others contribute more in creating the profits. Identifying these types of customers and their behavioural patterns, can facilitates the development of insightful business marketing strategies.

Since the UK and other western countries were put into national lockdown, eCommerce activity and sales have soared on websites like Amazon and Shopify. Shopify's total revenue for the full year of 2020 was more than $2.9billion, an 86% increase compared to 2019 *(Siliconrepublic)*. Moreover, Amazon recently reported $108.5 billion in sales in the first three months of 2021 which is up 44% from a year earlier *(Bussinesswire)*, meaning the pandemic has clearly driven online wholesale and retail sales even as countries 'open up'.
Ecommerce platforms will be analysing their customer behaviours through insightful techniques like Customer Life-time Value (CLV), Customer Retention (CR) and Customer Segmentation (CS) more than ever so they capitalise on the opportunity that the pandemic has provided.

CLV is a monetary value that represents the amount of revenue or profit a customer will give the company over the period of the relationship. CLV demonstrates the implications of acquiring long-term customers in contrast to short-term customers. For example, if customer 'A' is forecasted to spend $5,000 in the next two years, whereas customer 'B', is to only spend $1,000, subsequently eCommerce platforms will focus on certain customers and decide a marketing strategy. That being to retain the incoming cash flow from customer 'A' or a strategy to increase the incoming cash flow from customer 'B' (all on the basis of their predicted future revenue).

CLV helps eCommerce businesses by:

- Defining objectives for the company growth, expenditures, future sales and profits, etc.
- Optimising business marketing strategies,
- Adjusting campaigns and advertisement,
- Deciding on which cross sell and up sell products to recommend according to customer's purchase,
- Judging acquisition costs and the cost of attracting customers.

1.2 Aims & Objectives

The aim of this project is to reliably and successfully forecast a customers projected expenditure throughout their life-time on an eCommerce platform using modern machine learning methods, in terms of deep neural networks, and compare the results with conventional algorithm models found in literature/previous studies. Along with investigating further on customer behaviour exhibited on the chosen dataset like customer retention and segmentation.

Involving the objectives associated in this project are being able to:

- Carry out extensive literature reviews and find relevant studies associated with the dataset,
- Attain a background on Deep Neural Network and explain the technicalities involved,
- Pre-process the data in Python and exhibit adequate initial visualisations of the dataset,
- Generate useful business marketing insights in term of:
  Identify the most profitable customers,
  Suggest best product offerings to likely buying customers,

Distinguish active customers from inactive customers, forecasting transactions for individual customers,

Predict the purchase volume of the entire customer base,

Potential model customer segmentation on the profitable customers,

Possible customer retention modelling,

- Potentially construct a shortened version of this project analysis in a R shiny type format, for an easy overview read.

1.3 Inherent Challenges

Difficulties lie with the only data available is across 25 months and not a desirable four to five years, because this would provide more accurate life-time predictions. Typically, eCommerce datasets are proprietary and consequently hard to find among publicly available data, thus there has been extreme difficulty finding relevant datasets with a time span longer than 24 months. This relatively short time span of data may exhibit problems for the deep neural networks being used in this project to forecast customer future spending.
A study (linked to the same dataset with a similar aims) was found where they used Deep Neural Network (DNN). There were proposals given for future work on the DNN model which include, fine tuning the existing DNN model, involving more engagement features for the DNN model, etc.

A change may be made to the title of this project by incorporating Customer Retention (CR) and Customer Segmentation (CS). This is because this study will involve those aspects anyway and it may to necessary to contain deeper analysis on CR and CS and compare the results with conventional techniques from other studies. This will also widen the scope of the report.

2 Preliminary Literature Review

2.1 Highly Relevant Papers

Doing an initial literature review, valuable information has been discovered from certain online sources.

Starting with a study on RPubs, titled *Customer Lifetime Value for an Online Retail Store*. The report has aims very similar to this project and involves modern modelling techniques like deep neural network on the predicted sales. With a basic data visualisation at the beginning, the paper has useful insights into individual level estimates, cumulative actual sales compared to predicted transactions, transactional forecasts, DNN graph results with invoice dates and predicted sales. This study also offers recommendations on improving the DNN, which will be looked at in this report. *(RPubs)*.

The study titled, *Analysis of a public online retail dataset*, held a broad aim as purpose was to see what value could be extracted from the same dataset chosen for this project. The study's analysis includes some basic feature engineering and machine learning. It entails sections like Recency Frequency Monetary Value (RFM) and trend analysis, clustering, machine learning using conventional models on predicting customer spending. There is also a large section on suggested future work that could develop the forecasting results. *(GitHub)*

Next there is a study that portrays extensive insights and visualisations from the chosen dataset of this project. It is titled, *Online retail data analysis using R*. Consists of data cleaning and exploratory analysis, traditional machine learning algorithms on Spark to uncover more complex customer behaviour patterns, like which products are frequently purchased together. This study does not offer any recommendations in furthering their findings, though an advantage of this study is the deep insights into the attributes of the dataset. *(RStudio Pubs)*

2.2 Moderately Relevant Papers

Here are some papers/studies useful to this project, nonetheless their aim differs slightly to this project's aim.

A paper titled, *Customer Segmentation*, involved performing customer segmentation on the same eCommerce transaction dataset chosen here. After dividing customers into segments, K-means clustering analysis is carried out, then multiple linear regression on attributes like predicting the number of orders placed by each customer. Finally, there is a model evaluation section using cross validation and an improved model, random forest, is introduced. However, the study concluded that the models were ineffective in predicting customer segmentation. *(GitHub, Customer Segmentation)*

A brief study implicating the advantage and insight that RFM analysis can have on the chosen dataset. This study used classification models like logistic regression, Kneighbours classifier and decision trees to predict the clusters of customers using RFM. The study claims that its clusters predicted by the classification models aligns with K-means clustering, thus are correct. A disadvantage is that this study was vague and offered no future work section, gaps are visible in terms of not utilising modern algorithm models. This study was titled *Online Shoppers Intension Dataset*. *(Kaggle)*

The final study related to examining customer loyalty through clustering and RFM analysis. It is titled, *Customer Segmentation using RFM Analysis*, operating with common data mining techniques like K-means clustering. The point of grouping customers in this report was so that a customised marketing campaign could be made for each group of customers. This study presents useful information on the levels of customer loyalty associated with an eCommerce dataset and thus can be useful in this project. There is no real conclusion within this study and there were no modern data mining or deep neural network evaluation completed either. (Kaggle, CS using RFM Analysis)

3 Methodology

Locating a substantial and real eCommerce transactional dataset was not easy, this is because of the sensitive information that it holds about a company's business dealings. Nonetheless the chosen dataset was found on *(UCI Machine Learning Repository)* and titled *Online Retail II dataset*. This contains all the transactions occurring for a UK-based online retail store, with data between 1st December 2009 to 9th December 2011. This is an old dataset (that was made available in 2019 only), it is not the desired time span of four to five years, though as mentioned above finding a genuine eCommerce dataset of that time span was very difficult. The eCommerce company mainly sells unique all-occasion gifts, along with many of their customers being wholesalers.

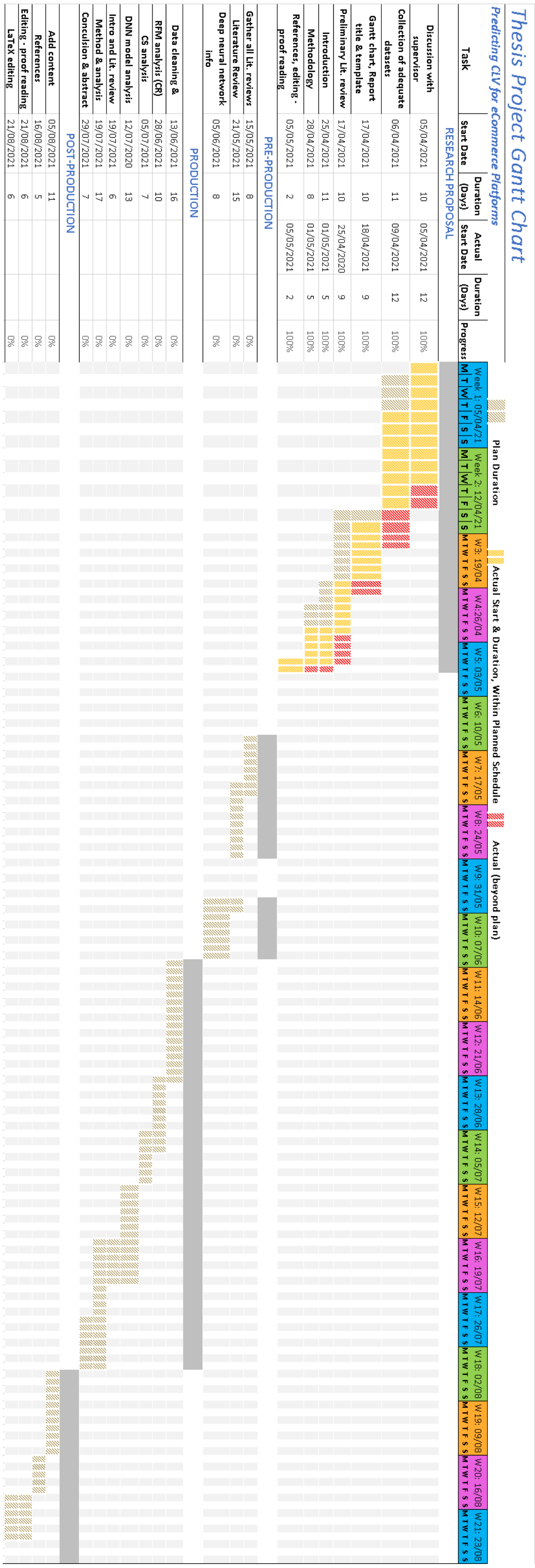
The dataset was kindly made available for public use by Dr Daqing Chen, Director of Public Analytics Group. See the acknowledgements section at the end of the References chapter to view further details. This project will be undertaken in Python and LaTeX, no further software or hardware is needed. It will contain firstly, an extensive pre-processing and visualisation presentation of the dataset offering what aspects need or possess evaluation ability.

To answer the aim of the project, customer life-time value will be calculated in connection with similar methods used in the found literature, then the forecasting portion will be attempted through deep neural networks to improve the existing conventional model methods. Comparisons will be made against the new models created in this project and whether they are worthy to replace existing models. The future work sections offered in most of the found literature will also be carried out to advance the ability to predict customer behaviour and expenditure.

4 Project Plan

4.1 Gantt Chart

Displaying this project's schedule with current schedule status. Deadlines of sections of work illustrated also. Please zoom in to see details



Thesis Project Gantt Chart

Predicting CLV for eCommerce Platforms

Task	Start Date	Duration (Days)	Actual Start Date	Duration (Days)	Progress
RESEARCH PROPOSAL					
Discussion with supervisor	05/04/2021	10	05/04/2021	12	100%
Collection of adequate datasets	06/04/2021	11	09/04/2021	12	100%
Gantt chart, Report title & template	17/04/2021	10	18/04/2021	9	100%
Preliminary Lit. review	17/04/2021	10	25/04/2020	9	100%
Introduction	25/04/2021	11	01/05/2021	5	100%
Methodology	28/04/2021	8	01/05/2021	5	100%
References, editing - proof reading	05/05/2021	2	05/05/2021	2	100%
PRE-PRODUCTION					
Gather all Lit. reviews	15/05/2021	8			0%
Literature Review	21/05/2021	15			0%
Deep neural network info	05/06/2021	8			0%
PRODUCTION					
Data cleaning &	13/06/2021	16			0%
RFM analysis (CR)	28/06/2021	10			0%
CS analysis	05/07/2021	7			0%
DNN model analysis	12/07/2020	13			0%
Intro and Lit. review	19/07/2021	6			0%
Method & analysis	19/07/2021	17			0%
Conculsion & abstract	29/07/2021	7			0%
POST-PRODUCTION					
Add content	05/08/2021	11			0%
References	16/08/2021	5			0%
Editing - proof reading	21/08/2021	6			0%
LaTeX editing	21/08/2021	6			0%

5 References

Siliconrepublic, *'Shopify revenue soars as Covid-19 boosts online shopping'*
https://www.siliconrepublic.com/companies/shopify-e-commerce-online-shopping
Accessed on: 3rd May 2021

Bussineswire; *'Amazon.com Announces First Quarter Results'*
https://www.businesswire.com/news/home/20210429006037/en/Amazon.com-Announces-First-Quarter-Results
Accessed on: 3rd May 2021

Literature

RPubs, *'Customer Lifetime Value for an Online Retail Store'*
https://rpubs.com/ragav208/CLV analysis
Accessed on: 2nd May 2021

GitHub, *'Analysis of a public online retail dataset'*
https://github.com/scheckley/online-retail
Accessed on: 2nd May 2021

RStudio Pubs, *'Online retail data analysis using R'*
https://rstudio-pubs-static.s3.amazonaws.com/430563d38c12b53d724fa6852949b1f3e4ffbf.html
Accessed on: 2nd May 2021

GitHub, *'Customer Segmentation'*
https://gioiacopini.github.io/customer.html
Accessed on: 2nd May 2021

Kaggle, *'Online Shoppers Intension Dataset'*
https://www.kaggle.com/surekharamireddy/e-commerce-data-set/notebook
Accessed on: 2nd May 2021

Kaggle, *'Customer Segmentation using RFM Analysis'*
https://www.kaggle.com/rajpraveenpradhan/customer-segmentation-using-rfm-analysis
Accessed on: 2nd May 2021

UCI Machine Learning Repository, *'Online Retail II Data Set'*
https://archive.ics.uci.edu/ml/datasets/Online+Retail+II
Accessed on: 19th April 2021

Acknowledgements

For the UCI Machine Learning Repository, this data was made available by Dr Daqing Chen, Director of Public Analytics Group. chend '@' lsbu.ac.uk, School of Engineering, London South Bank University, London SE1 0AA, UK.